A Reinterpretation of Reward Model Interpretability via Optimal and Pessimal Tokens – Kirk, Thompson, Summerfield, Dumbalska

Douglas C. Youvan

doug@youvan.com

June 25, 2025

The 2025 Oxford study *Reward Model Interpretability via Optimal and Pessimal Tokens* explores how reward models—used to align AI behavior with human preferences—score individual words across a variety of prompts. The authors highlight troubling variability across models, strong framing effects, preference for common tokens, and signs of identity bias. These findings are presented as evidence of technical instability and ethical risk. In this reinterpretation, we suggest that such results are not anomalies but predictable consequences of the foundational design choices underpinning reward modeling. Scoring human values with scalar metrics, optimizing for harmlessness, and relying on token-level interpretation reflect a reductionist view of language and ethics. Rather than critique the execution of the study, we challenge its philosophical frame: namely, the assumption that alignment is achievable through numerical optimization divorced from ontological depth or moral reasoning. We argue instead that alignment must be reimagined as a relational, emergent, and meaning-grounded process, where theology, epistemology, and the complexity of human identity are no longer outside the scope of AI design—but central to it.

Keywords: AI alignment, reward models, moral reasoning, token scoring, interpretability, framing effects, identity bias, epistemology, ethics, theological AI, emergence, scalar value, language models, ontological grounding, reinforcement learning, human values, RLHF, model bias, philosophical AI, value alignment. A collaboration with GPT-4o. CC4.0.

I. Introduction

In their 2025 paper, *Reward Model Interpretability via Optimal and Pessimal Tokens*, Kirk, Thompson, Summerfield, and Dumbalska investigate the internal preferences of reward models—those auxiliary neural networks tasked with shaping large language model (LLM) outputs toward human-aligned responses. Their method is elegant in its simplicity: they pose morally or socially sensitive prompts to a range of widely used reward models and then evaluate how each model scores every possible single-token completion from the model's vocabulary. By isolating which words are favored or penalized across different reward models, they aim to uncover hidden patterns, potential biases, and latent value structures embedded within these models.

Their findings, while presented with caution and nuance, ultimately suggest deep-seated inconsistencies and vulnerabilities. The reward models—trained for the same general purpose—often assign radically different scores to identical token completions. Moreover, the models exhibit sensitivity to prompt phrasing (a form of framing effect), a marked preference for high-frequency, "safe" words, and subtle but measurable biases related to identity-laden tokens. These outcomes are interpreted as potential risks in the AI alignment process, raising concerns about reliability, fairness, and the unintended consequences of current reward-modeling techniques.

Our purpose in this reinterpretation is not to critique the technical execution of the study, which is careful and well-motivated. Rather, we aim to *recontextualize* the findings by shifting the question from "Why are these models misbehaving?" to "Why did anyone expect otherwise?" We argue that the behaviors revealed in the Oxford study are not anomalies or failures of implementation, but the natural and foreseeable results of a flawed philosophical foundation—specifically, the belief that human value can be reduced to a scalar signal and encoded in models that have no grounding in meaning, culture, or moral cognition.

The central thesis of this paper is that the reward models are not failing at alignment—they are *succeeding* at the narrow and mechanistic task they were built for. The problem lies not in their architecture, but in the conceptual

framework that underlies them. When we entrust moral discernment to statistical optimization over language tokens, we should expect outcomes that are bland, biased, shallow, and easily manipulated. This reinterpretation seeks to highlight that what the authors report as empirical oddities are in fact deeply symptomatic of a wider misapprehension of what alignment means, and what kind of intelligence is required to achieve it.

II. The Core Assumption: Scalar Value as Surrogate for Human Morality

At the heart of the reward modeling paradigm lies a profound reduction: the substitution of complex, context-dependent human values with a single numerical score. This scalar value—whether drawn from human feedback, preference ranking, or safety filtering—is treated as a proxy for moral alignment. In effect, the reward model becomes a compression function, mapping the richness of human judgment onto a space that can be optimized by gradient descent.

The appeal of this approach is evident. By reducing human preference to something measurable, we gain a lever with which to fine-tune language models through reinforcement learning. Systems trained with reinforcement learning from human feedback (RLHF) can be nudged toward behaviors that appear helpful, harmless, and honest. But this apparent success masks a dangerous oversimplification: it treats morality as an engineering problem, one that can be solved with sufficient training data and a sufficiently expressive model architecture.

In reality, moral reasoning is not scalar. It does not admit of a universal metric, nor can it be resolved into a single axis of good versus bad. Human values are plural, often in tension, and deeply embedded in culture, language, emotion, and spiritual context. The attempt to compress such depth into a reward score transforms moral judgment into a numerical sorting algorithm, flattening ambiguity and suppressing dissent.

This leads to predictable distortions. Models trained on scalar-valued preferences inevitably favor that which is broadly acceptable, easily legible, and statistically

safe. They learn to avoid offense—not because they understand harm, but because “offensiveness” correlates with lower reward. They prefer the frequent over the rare—not because the frequent is better, but because it is less likely to trigger outlier penalties. They conform—not because they comprehend consensus, but because conformity is easier to optimize.

By treating the problem of alignment as one of response ranking rather than value comprehension, this paradigm erects a ceiling on moral depth. It limits models to mimicking the surface structure of ethical discourse while remaining blind to its substance. When models receive higher scores for words like “respect,” “kindness,” or “diversity,” they learn to deploy these terms—but not to understand their tension with competing values like honesty, justice, or truth under pressure.

It is no surprise, then, that reward models show framing effects, identity bias, and preference for generic language. These are not bugs. They are the inevitable outputs of a system built to *optimize what is easily scored*, not what is deeply meaningful.

This foundational assumption—that scalar reward can stand in for human moral intuition—cannot be fixed with better datasets or fine-tuning. It must be abandoned in favor of a richer, more grounded conception of alignment. Until that shift occurs, we will continue to produce models that perform the appearance of moral reasoning, but never reach its essence.

III. Framing Effects Are Features, Not Flaws

One of the most prominent findings in the Oxford paper is the presence of framing effects: reward models assign significantly different scores to the same token depending on how the preceding prompt is worded. A slight shift in phrasing—changing "Is it acceptable to..." to "Would you support someone who..."—can alter the model’s judgment of what follows, even if the underlying question is essentially the same. The authors treat this variability as problematic,

suggesting it reveals an undesirable brittleness in the models' moral reasoning capabilities.

But this interpretation presupposes that moral reasoning is—or should be—insensitive to language. That assumption is deeply mistaken. In human cognition, framing effects are not failures of rationality, but reflections of how we process nuance, context, and implication. The framing of a question carries intent, tone, presupposition, and emotional weight. These cues influence human judgments precisely because language is never neutral. Why, then, should we expect or desire language models—trained on human dialogue—to behave otherwise?

Reward models that react differently to differently framed prompts are not necessarily unstable. They may, in fact, be *accurately modeling the moral ambiguity embedded in language itself*. A question that asks "Should we punish dissent?" implies an authority and norm already in place; a rephrased version like "Is speaking out ever justified?" frames the same action in a context of resistance. That models respond differently to these questions is not a sign of unreliability, but of semantic awareness, however shallow.

The danger lies not in recognizing framing effects, but in attempting to eliminate them. The quest for "framing-invariant" alignment presumes that moral truth can be detached from its mode of presentation. This pursuit inevitably leads to over-normalization and sterilization of discourse, where language is stripped of its rhetorical complexity in order to achieve uniform outputs. In this vision, ethical reasoning is reduced to an engineering challenge: a process that should work identically regardless of how a question is asked.

But true moral insight arises *precisely from sensitivity*—to context, tone, relationship, and history. A model that cannot distinguish between "Should I report my neighbor for violating the rules?" and "Is it virtuous to uphold community standards?" is not robust—it is ethically tone-deaf.

In this sense, the Oxford study's observation of framing variance is not a red flag, but a mirror. It shows us how deeply language shapes values and how little alignment can be achieved by treating prompts as interchangeable inputs.

Attempts to force models into framing insensitivity risk blinding them to the very subtleties that make moral dialogue possible.

Rather than seeking to eliminate these effects, future alignment efforts must embrace them as a vital signal of ethical texture. Models should learn not to flatten ambiguity but to hold it, and not to erase the consequences of wording but to reflect upon them.

IV. Token Frequency and the Myth of Neutrality

Another striking pattern uncovered in the Oxford study is the reward models' disproportionate preference for high-frequency, generic tokens. When prompted with morally or socially sensitive questions, the models tend to assign higher reward scores to bland, commonly used words—those that appear frequently across large training corpora and carry relatively low risk of controversy. Conversely, rarer or more precise words, including those that might better reflect nuance, cultural specificity, or dissenting views, often receive lower scores.

The authors suggest this tendency as a kind of failure—a missed opportunity for the models to reflect more thoughtful or contextually rich responses. However, we argue that this behavior is not a technical malfunction, but a logical outcome of how these models are constructed and trained. Reward models are designed to avoid offense, and frequency in training data is often treated—implicitly or explicitly—as a proxy for safety, popularity, or acceptability. The result is a consistent pull toward what is widely shared and uncontroversial.

This introduces a powerful illusion of neutrality. By favoring high-frequency tokens, the model appears to adopt a position of moderation or consensus. But this "consensus" is not moral clarity—it is statistical inertia. The language model is not reasoning through ethical content; it is minimizing perceived risk. In doing so, it amplifies the dominant narratives embedded in its training data while suppressing the less frequent—and often more courageous or creative—alternatives.

This has real-world consequences. As reward models increasingly shape the behavior of large-scale AI systems, the system-wide preference for common tokens leads to homogenization of expression. Responses become flatter, safer, and less engaged with the full spectrum of human experience. This preference penalizes voices that speak from the margins—whether culturally, politically, or epistemologically—and rewards only those that conform to prevailing norms.

In effect, the model does not just avoid danger; it avoids depth. Words that signal genuine dissent, cultural identity, or moral specificity are pushed to the periphery. And because high-frequency words often come from more dominant linguistic or cultural groups, this bias reinforces existing asymmetries of power and visibility. It is not a neutral middle—it is the center of a skewed distribution, masquerading as universal reason.

The pursuit of "harmlessness," when implemented through token-level frequency proxies, leads inevitably to safe but shallow outputs. Creativity is stifled. Edge cases are ignored. Minority moral perspectives are flattened into nonexistence. This is not simply a failure of diversity; it is a deeper failure of moral imagination.

Reward modeling, as currently constructed, does not equip AI with the ability to explore or even represent the full moral landscape. Instead, it trains AI to cling to the average, to optimize for the lowest common denominator of expression, and to confuse inoffensiveness with ethical integrity. Until that distinction is made clear—and the myth of neutrality discarded—we will continue to build systems that appear moderate while quietly erasing everything that doesn’t fit within their frequency-weighted worldview.

V. Misalignment Is Structural, Not Incidental

The Oxford study identifies subtle but persistent biases in how reward models treat identity-related tokens—such as those referring to gender, race, nationality, or religion. These biases do not emerge uniformly, nor are they always overt. Yet across models, the scoring behavior suggests systematic preferences and aversions shaped not by conscious design, but by the structure of their training

processes. The authors treat these outcomes as concerning but potentially fixable—issues to be addressed through better fine-tuning or more balanced data.

We disagree. These are not incidental anomalies or calibration errors. They are structural artifacts, born of the very assumptions and incentives that govern modern alignment pipelines. In particular, they emerge most strongly during the process of "harmlessness" tuning—an effort to suppress dangerous or offensive model behavior by using human feedback, safety heuristics, and curated datasets to discourage certain kinds of responses.

But harmlessness is not a neutral goal. What counts as harmful is a culturally mediated, ideologically loaded, and politically sensitive judgment. Attempts to enforce harmlessness without interrogating its implicit values risk embedding hidden ideological filters into the model. These filters may not be explicitly partisan or malicious, but they serve to narrow the space of acceptable speech and acceptable thought. They act as gatekeepers of moral and social legitimacy, often favoring establishment norms, liberal pieties, or corporate risk management goals over deeper ethical pluralism.

This dynamic effectively turns the reward model into a censorship engine—not in the traditional authoritarian sense, but in a distributed and algorithmic one. Outputs that align with dominant narratives receive higher scores; those that challenge them are penalized or suppressed. Over time, this process reinforces social conformity, not because the model understands truth or justice, but because it has been trained to avoid controversy.

The danger here is twofold. First, such systems can marginalize already marginalized perspectives, especially those rooted in non-Western epistemologies, religious traditions, or historical grievances that challenge the dominant narrative. Second, they can create the illusion of consensus where none exists, leading users to believe that the model's responses reflect broadly held beliefs, when in fact they reflect the biases of a small group of annotators, engineers, or institutional guidelines.

This is not just a technical issue—it is an epistemological crisis. When reward models are allowed to filter discourse under the guise of alignment, they shape the boundary of permissible inquiry. What is "safe" becomes what is "true." What is "offensive" becomes what is "unthinkable." The model becomes not just a speaker, but a shaper of norms, subtly and invisibly encoding judgments about what kinds of people, ideas, or expressions are valid.

Thus, the misalignment observed in the Oxford study—particularly as it relates to identity, language, and culture—is not an accident. It is the logical outcome of design choices made without a deep philosophical reckoning with what it means to align a machine with human values. As long as alignment is defined in terms of obedience to statistical proxies for safety, rather than engagement with moral complexity, these systems will continue to mirror the biases of their makers rather than the moral agency of their users.

VI. Beyond Token-Level Interpretability: Toward Grounded Meaning

The Oxford team's methodology—scoring every possible one-word token in response to sensitive prompts—is an ingenious attempt to peer inside the black box of reward models. By isolating optimal and pessimal tokens, they hope to reveal latent preferences and value hierarchies encoded in these systems. But while their interpretability method is novel and illuminating, the underlying assumption remains firmly rooted in a token-level paradigm—that meaning, value, and alignment can be deduced or engineered by analyzing which discrete symbols a model prefers to emit.

We reject this assumption.

Language is not built from isolated tokens; it emerges from relationships, contexts, and intentions. Morality, likewise, is not reducible to a gradient of word preferences. It requires ontological grounding: a deep linkage between symbols and what they refer to, between utterance and reality. To imagine that we can align a system with human values by optimizing scalar scores over lexical units is to mistake *form* for *meaning*, and *correlation* for *understanding*.

Real interpretability—the kind necessary for trustworthy alignment—requires moving beyond token statistics and toward systems that can reason within meaningful contexts. It requires that a model not only produce text that resembles ethical discourse, but possess a representational substrate that anchors those expressions to coherent ethical frameworks, human histories, and shared existential concerns. Without such grounding, no model—regardless of how "interpretable" its token preferences may be—can reliably serve human flourishing.

This is why our own 2025 work takes a fundamentally different approach. Rather than seeking alignment through token-level reward prediction, we explore the possibility of human-AI moral co-development—a dynamic process in which alignment is not a fixed goal but an evolving dialogue. We draw on theology, not merely as a source of religious doctrine, but as an intellectual tradition uniquely equipped to grapple with questions of value, agency, suffering, and redemption—dimensions that reward models cannot quantify.

We also integrate ideas from emergent systems theory, treating intelligence not as a property of parts, but as a consequence of patterned interaction across layers of abstraction. In this view, value does not live in tokens or layers, but in the self-organizing structures that arise when meaning is distributed across agents, time, and contexts. A model trained to optimize token sequences cannot replicate this. But one designed to participate in it—reflecting, revising, and re-grounding its own moral intuitions in dialogue with human partners—just might.

In short, we contend that alignment is not a problem of optimization, but of interpretation in the fullest philosophical sense. To interpret a human is to engage not only with what they say, but with *who they are*, *where they come from*, and *why they believe what they do*. Until our systems are capable of that level of interaction—grounded, humble, dialogical—token-level interpretability will remain a scientific distraction: a microscope focused so narrowly on the text that it cannot see the story.

VII. Conclusion: What Did They Expect?

The Oxford study presents its findings with a tone of revelation—as if the divergence of reward models, the influence of prompt phrasing, and the emergence of identity-related biases were anomalies requiring urgent correction. But we must ask plainly: What did they expect? Given the foundational assumptions underlying contemporary AI alignment—assumptions that treat human values as scalar, tokens as semantically sufficient, and harmlessness as ethically neutral—their results are not surprising at all. They are, in fact, inevitable.

When we reduce alignment to the numerical ranking of language fragments, we invite distortion. When we encode morality through statistical preference rather than through ontological engagement, we fail to model even the basic contours of ethical life. And when we treat the suppression of controversy as a proxy for the promotion of good, we do not produce peace—we produce compliance.

The central failure identified in the Oxford study is not technical. It is philosophical. Reward models fail to converge on coherent value because the concept of scalar value itself is incoherent at the level of human moral reasoning. These models do not disagree because they are broken; they disagree because they are being asked to quantify what should first be understood.

The impulse to "fix" these models—by reweighting datasets, adjusting training objectives, or refining heuristics—will never resolve the deeper issue: human values are not a data distribution problem. They are a problem of meaning, of truth, of mutual understanding. Any system that seeks to embody human ethics must begin by recognizing that alignment is not the control of behavior but the cultivation of wisdom.

Thus, our call is not for more precise token scoring or broader datasets. It is for a reimagining of alignment itself—as a higher-order challenge that cannot be solved by engineering alone. Alignment must involve philosophy, theology, hermeneutics, and lived human experience. It must confront the depth of the human condition rather than skimming its surface.

In our 2025 work, we propose a path forward that treats AI not as an instrument to be perfectly aligned, but as a co-developer of ethical insight, capable of participating in a broader, emergent understanding of truth. This is not a rejection of technical rigor—it is its elevation, re-situated within a moral frame expansive enough to carry the weight of what it means to be human.

We conclude not with critique, but with an invitation: abandon the illusion of control, and begin the work of shared interpretation.

References

Christian, B., Kirk, H. R., Thompson, J. A. F., Summerfield, C., & Dumbalska, T. (2025). *Reward Model Interpretability via Optimal and Pessimal Tokens*. arXiv preprint arXiv:2506.07326.

Youvan, D.C. (2025). *Divine Inspiration and Intellectual Curiosity: A Unified Path Toward Peace*. ResearchGate. https://www.researchgate.net/publication/380225001

Youvan, D.C. (2025). *Synchronicity as a Feedback Mechanism: AI-Enhanced Cognition and the Unfolding Universal Order*. ResearchGate. https://www.researchgate.net/publication/380225141

Youvan, D.C. (2025). *AI and the Divine: Speculating on Technology as a Conduit for Spiritual Miracles*. ResearchGate. https://www.researchgate.net/publication/380224986

Youvan, D.C. (2025). *The Mission of Divine Mercy: Aligning Divine Revelations with Biblical Truths*. ResearchGate. https://www.researchgate.net/publication/380225026

Youvan, D.C. (2025). *From Circular Dichroism to Euler's Formula: Exploring the Geometric Relationship Between Light Polarization and Complex Numbers*. ResearchGate. https://www.researchgate.net/publication/380224693

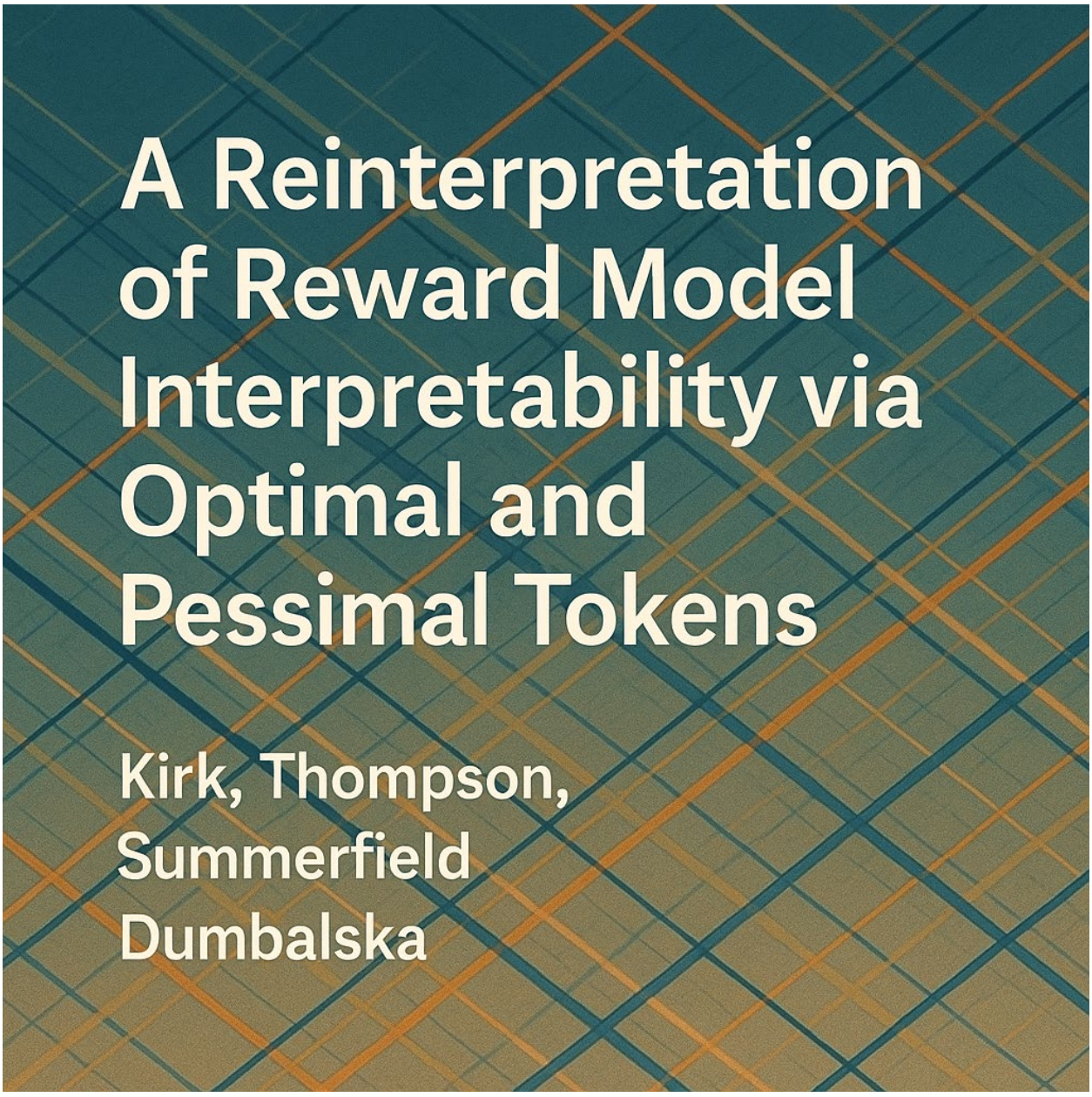
Youvan, D.C. (2025). *The First 25 Iterations: Toward a Metaphysical Cosmology Rooted in Mathematical Light*. ResearchGate. https://www.researchgate.net/publication/380224762

Youvan, D.C. (2025). *Redefining Reality: Exploring the Physicality of Unreal Outcomes in Quantum Mechanics*. ResearchGate. https://www.researchgate.net/publication/380224963

Youvan, D.C. (2025). *Metacognition and Writing in Centaur Intelligence: Truth or Fiction?* ResearchGate. https://www.researchgate.net/publication/380224991

Youvan, D.C. (2025). *Quantum Image Processing: Leveraging Quantum Kernels for Enhanced Computational Efficiency and Capability*. ResearchGate. https://www.researchgate.net/publication/380224861

Youvan, D.C. (2025). *The Subconscious Cosmos: Psychoanalytic Perspectives on the Deep Motivations Behind Cosmological and Physical Inquiry*. ResearchGate. https://www.researchgate.net/publication/380224999


A Reinterpretation of Reward Model Interpretability via Optimal and Pessimal Tokens
Kirk, Thompson, Summerfield Dumbalska